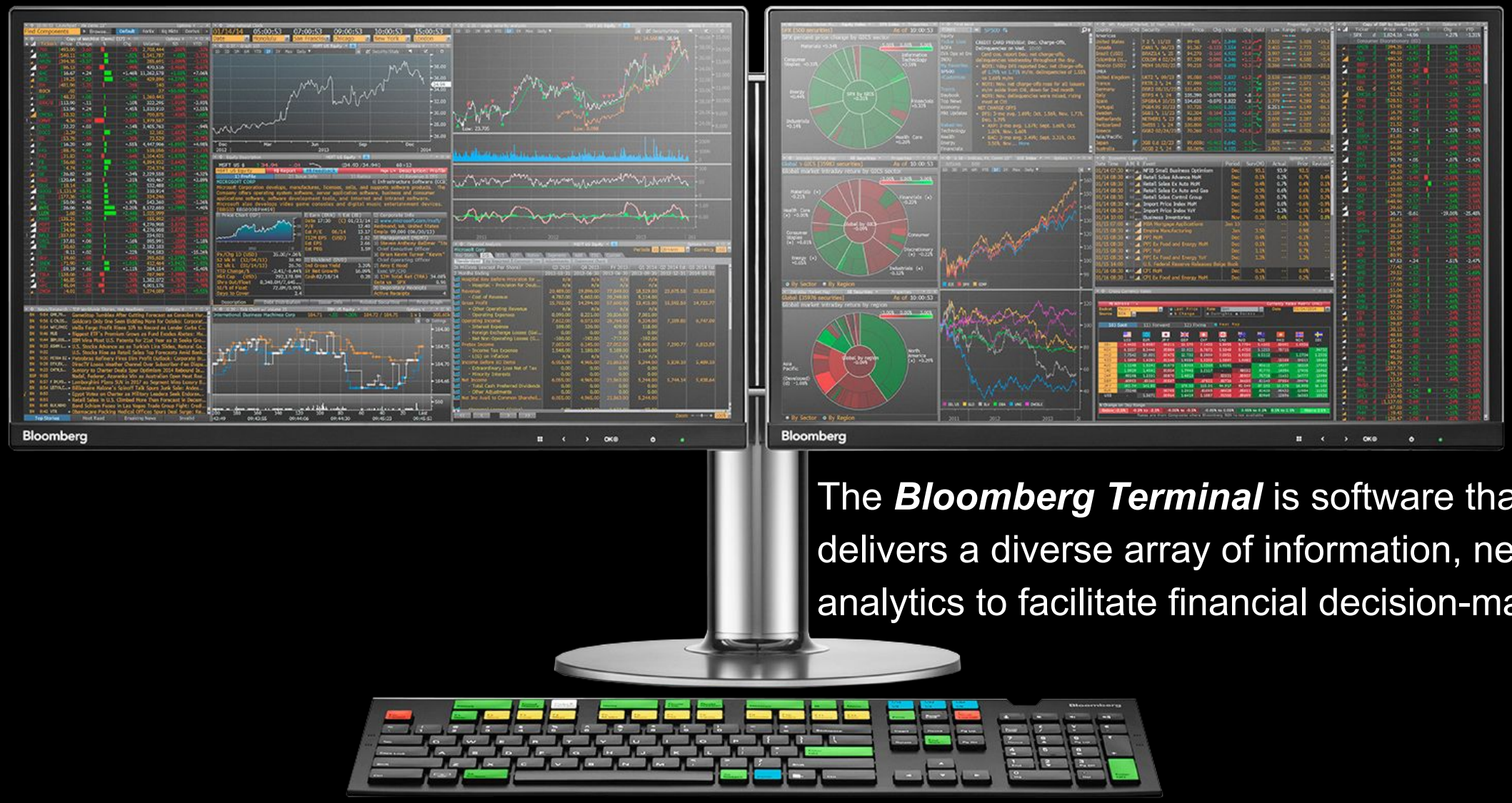
Applied Named Entity Recognition @ Bloomberg

Bloomberg Engineering

NLP 280: Seminar in Natural Language Processing
UC Santa Cruz
November 6, 2020

Daniel Preoțiuc-Pietro
Senior Research Scientist, AI Group

TechAtBloomberg.com



The ***Bloomberg Terminal*** is software that delivers a diverse array of information, news and analytics to facilitate financial decision-making.

TechAtBloomberg.com

Bloomberg

Engineering

Bloomberg is just finance, right?

A technology company founded in 1981

Our strength and focus are data, analytics, and community tools

Both creator and consumer of news

~20,000 employees in 167 offices around the globe

- More than 2,700 journalists and analysts
- 6,000+ software engineers, including more than 150 engineers and data scientists working on AI problems

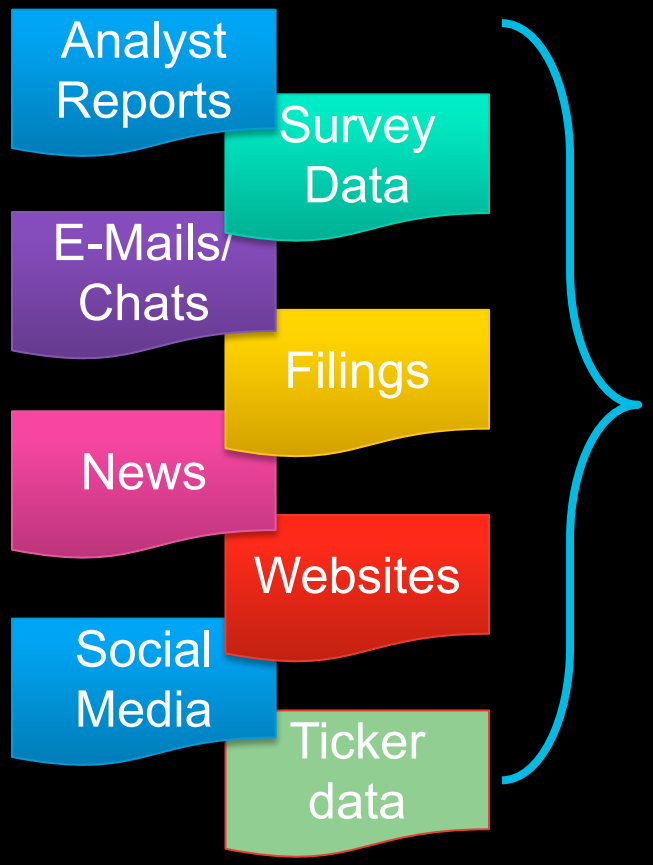
Increased use of and contributions to open source software

TechAtBloomberg.com

Bloomberg

Engineering

From Data to Client Applications



Bloomberg

Bloomberg LAW

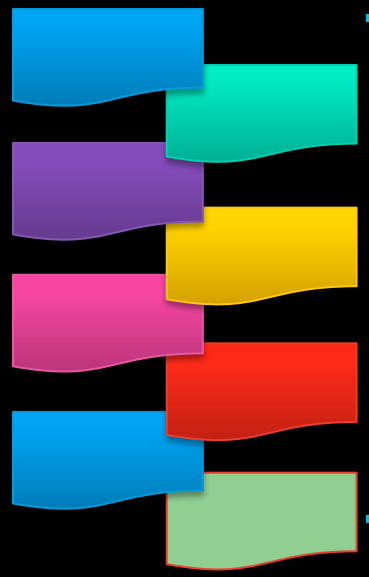
Bloomberg GOV

TechAtBloomberg.com

Bloomberg Engineering

AI at Bloomberg

EDITORIAL



Bloomberg

COMMUNICATIONS

NEWS

INCUBATION + FINANCE

Bloomberg LAW

Bloomberg GOV

DISCOVERY

SEARCH

PLATFORMS

TechAtBloomberg.com

150+ Research Engineers and growing!

Bloomberg

Engineering

Named Entities in Finance

Named entities are central to financial documents

- 80% of financial data remains unstructured
- Named entities are essential to understanding unstructured content

TechAtBloomberg.com

Bloomberg

Engineering

Entity Detection - Entity Recognition

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

NER

ORG PER TIC ORG

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

Goal: Identify named entities in text (offsets) and their types

Bloomberg

Engineering

Entity Detection - Entity Disambiguation

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

NER

ORG PER TIC ORG

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

NED

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

AMZN US 216952 BIO 20099335 AMZN US 216952 HRP

Goal: Associate named entity mentions in a text to their unique identifier in a knowledge base

Bloomberg

Engineering

Entity Detection - Entity Salience

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

NER

ORG PER TIC ORG

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

NED

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

AMZN US 216952 BIO 20099335 AMZN US 216952 HRP

Salience

AMAZON SUSPENDS TOP ENTERTAINMENT CHIEF Price in Wake of Harassment Claim $AMZN – Hollywood Reporter

AMZN US 216952 BIO 20099335 AMZN US 216952

Goal: Quantify how central is each explicitly mentioned named entity to the text

Bloomberg Engineering

Entity Detection in Practice

Defiant Johnson Meets Irish Leader for Talks: Brexit Update

By Alex Morales and Kitty Donaldson

(Bloomberg) -- The beleaguered U.K. Prime Minister Boris Johnson is in Dublin on Monday for talks with his Irish counterpart, Leo Varadkar, as he presses ahead with his hardline plan to leave the European Union "do or die" by Oct. 31.

Key Developments:

- Irish Finance Minister Paschal Donohoe says his country is open to Brexit extension
- Parliament set to vote again on an early general election on Monday evening, with opposition parties expected to reject the measure
- Over the weekend, Amber Rudd quit the cabinet with a furious attack on Johnson's leadership
- Chancellor of the Exchequer Sajid Javid and Foreign Secretary Dominic Raab said on Sunday that the Brexit plan is unchanged

1) Kitty Donaldson (Bloomberg LP)
2) Thomas Penny (Bloomberg LP)
3) Leo Varadkar (Republic of Ireland)
4) Theresa May (United Kingdom of Great Britain and Northern...
5) Jeremy Bernard Corbyn (United Kingdom of Great Britain an...
6) Dominic Raab (United Kingdom of Great Britain and Norther...
7) Paschal Donohoe (Republic of Ireland)
8) Amber Augusta Rudd (United Kingdom of Great Britain and ...
9) Tim Ross (Bloomberg LP)
10) Sajid Javid (United Kingdom of Great Britain and Northern I...
11) Boris Johnson (United Kingdom of Great Britain and Norther...
12) Dara Doyle (Bloomberg LP)
13) Alex Morales (Bloomberg LP)

Bloomberg

Engineering

Applications at Bloomberg

Entity detection is technology useful across the company

- News and financial document accessibility
- News editing
- Search and autocomplete
- ETL pipelines
- Enhancing client communications

Entity detection is used in many Bloomberg applications

- Directly
- As part of a pipeline of models

Trends

SPX | Set Alert | Set View as Default | News Trends

Trend Period 8 Hours

News Reader Activity | News Sentiment | News Volume | Twitter Volume

Largest Increase | Largest Total

	Security	Δ Publication	GT	Δ Price	Δ AVAT	Sent.	Representative Tweet
1)	Kraft Heinz Co/The			-1.84%	-51.94%	-0.11	Wall Street Journal: The story of Kraft Hei...
2)	Citrix Systems Inc			+0.34%	+23.50%	0.00	$Stocknewsalerts: $CTXS New SEC Docume...
3)	Franklin Resources Inc			+1.18%	+1.27%	0.00	Daily Political: Franklin Resources Inc. Ha...
4)	Wells Fargo & Co			+0.28%	+7.09%	-1.00	CNBC Now: BREAKING: Wells Fargo says it ...
5)	Conagra Brands Inc			+1.49%	+31.02%	0.00	stockmarketnow: ConAgra CEO hopes bord...
6)	Bristol-Myers Squibb Co			+0.35%	+176.51%	0.00	Wall Street Journal: Carl Icahn has taken ...
7)	QUALCOMM Inc			+0.51%	-20.79%	+0.01	Engadget: Intel and Qualcomm are steadil...
8)	First Solar Inc			+5.11%	+115.90%	-0.75	Bloomberg: First Solar heads into a restru...
9)	Phillips 66			+0.47%	+27.69%	0.00	Phillips 66: Mayor by day, shift supervisor...
10)	Newmont Mining Corp			+1.16%	+2.61%	-0.19	Estimize: Well done @apeppecali beating 1...
11)	Urban Outfitters Inc			+0.62%	+0.11%	0.00	FOX40 News: Is This Retro or Sad? Urban ...
12)	FirstEnergy Corp			+1.03%	+17.73%	0.00	Andrew Scurria: FirstEnergy Corp. reports ...
13)	CarMax Inc			+1.84%	-30.43%	0.00	Zolmax News: Alyeska Investment Group ...
14)	Yahoo! Inc			+0.89%	-13.82%	-0.39	James Cook: RT @alexweprin: Here's the ...
15)	Microchip Technology Inc			+1.12%	+69.97%	0.00	WKRB News: Microchip Technology Incorpo..
16)	Home Depot Inc/The			+1.41%	+81.56%	+0.04	Benzinga.com: Kevin Kelly's Home Depot ...
17)	CSX Corp			+0.76%	-36.01%	0.00	The NOW Team RE/MAX: CSX announces th...
18)	Mead Johnson Nutrition Co			-0.03%	-24.54%	+0.11	Zolmax News: Alyeska Investment Group ...
19)	Bed Bath & Beyond Inc			+0.78%	+46.00%	0.00	InterCooler: 4,845 Shares in Bed Bath & B...
20)	Travelers Cos Inc/The			-0.33%	-6.68%	0.00	Watchlist News: The Travelers Companies,...
21)	Concho Resources Inc			+0.92%	+64.19%	0.00	Conf Call Tran: Concho Resources Inc. Rep...
22)	Visa Inc			+0.53%	+57.98%	0.00	Ophir Gottlieb: $V Shh... There's edge in ...

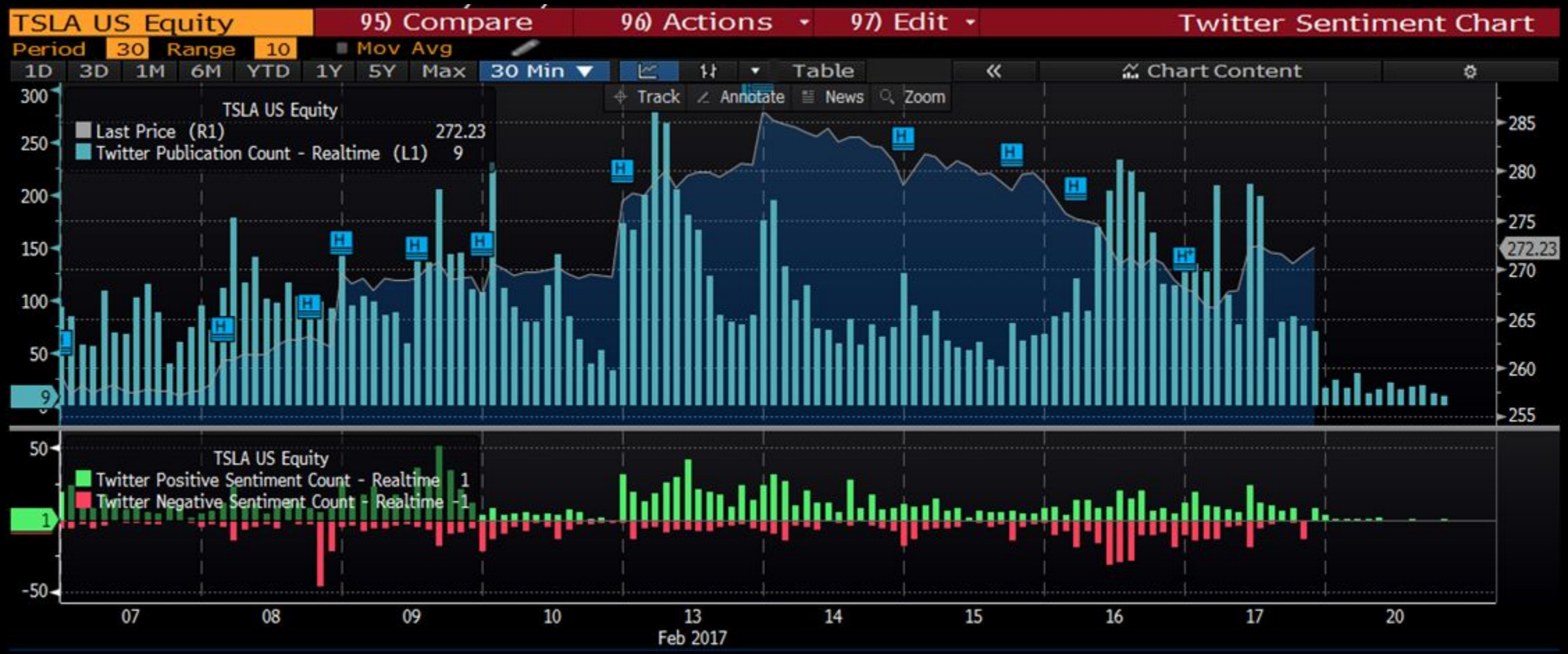
Entity Recognition, Entity Disambiguation and Entity Salience

TechAtBloomberg.com

Bloomberg

Engineering

Sentiment



Entity Recognition, Entity Disambiguation, Entity Salience and Targeted Sentiment Analysis

TechAtBloomberg.com

Bloomberg

Engineering

Company Alerts

All Companies — Bloomberg Social Velocity Monitor

Market Reaction — Stronger

		Time	Company Name	Equity Price	Equity %Chg	Equity Δ AVAT	Option Volm	Option Δ Volm	Snmt	News Headline
10)		11:21	GENERAL MOTORS C	30.99	-.64%	8.7%	13696	-57.5%		Bell Tolls for Car Owners in GM Ignition ...
11)		09:27	VERINT SYSTEMS	33.49	-4.83%	1891.6%	4176	1073.4%		VERINT STREET WRAP: Shares May Remai...
12)		09:39	FITBIT INC - A	13.40	-4.35%	80.5%	18184	-19.3%		Fitbit Sinks 7.7%, Continues to Underperf...
13)		09:35	BOEING CO/THE	128.58	-1.76%	27.1%	30190	50.7%		Boeing's Retooled Plane Unit Aims to Dul...
14)		12:16	DUKE ENERGY CORP	80.35	-.11%	-59.9%	707	-82.6%		CharlotteObservr: BRIEF: Duke Energy do...
15)		10:43	METLIFE INC	44.73	5.35%	247.3%	30755	318.5%		MetLife Gains Capital Flexibility as CEO E...
16)		10:55	GOLDMAN SACHS GP	156.50	.95%	-24.3%	21428	-.4%		Jefferies Said to Hire Goldman's Fernenia...
17)		16:27	MEDIVATION INC	37.39	-3.51%	43.9%	6176	76.9%		Medivation Said to Work With Advisers to ...
18)		09:46	HIMAX TECHNO-ADR	11.66	-2.83%	57.7%	12134	274.2%		Seeking Alpha: Himax's recent strength a ..
19)		16:20	PROGRESS SOFTWAR	25.59	1.11%	27.4%	885	1576.1%		Progress 2016 Rev. View Misses Ests., CF...
20)		10:55	NETFLIX INC	102.19	-1.86%	2.4%	129590	-8.3%		Q. and A.: Ashton Kutcher Talks About His...
21)		10:23	BANK OF AMERICA	13.48	.45%	-23.9%	210667	-29.0%		Bank of America Top Bond Salesman Brya...
22)		03/29	AMGEN INC	149.48	-.01%	-41.0%	5749	-46.7%		Fidelity Health Care Adds Alexion, Exits D...
23)		09:50	ALPHABET INC-C	750.53	.77%	-6.9%	34493	45.9%		Apple Wins Ruling Invalidating One Smart...
24)		15:39	SQUARE INC - A	15.02	9.32%	62.4%	5725	-8.7%		Square Gains as Much as 10%; Mizuho Call...
25)		14:29	GALENA BIOPHARMA	1.38	13.11%	168.7%	10419	58.4%		RXi Pharmaceuticals Reports Fourth Quart...
26)		03:50	TOSHIBA CORP	207.00	-2.13%	-27.8%	--	--		Toshiba Recalls Laptop Computer Battery ...
27)		09:53	JOHNSON&JOHNSON	108.98	-.15%	-34.2%	14875	-22.5%		Germab/J&J Multiple Myeloma Study Stop...
28)		09:23	KELLOGG CO	76.83	.31%	-3.8%	514	-60.8%		Seeking Alpha: Kellogg's: 'They're Grrrr...
29)		09:48	TWITTER INC	16.36	2.44%	-.2%	84035	4.3%		Turkey's Bozdag Asks Twitter Who Ordere...
30)		16:48	MICROSOFT CORP	55.05	.62%	-22.5%	114292	39.6%		Microsoft CEO Stays Committed to AI Bots...
31)		09:47	FACEBOOK INC-A	114.70	-1.24%	42.1%	363316	54.4%		Facebook Billionaires' Wealth Hits Record...
32)		15:30	RADIUS HEALTH IN	32.83	6.45%	164.1%	1046	8.0%		Radius Health Climbs After Filing Abalopa...
33)		16:21	ATYR PHARMA INC	4.86	-3.76%	34.5%	--	--		Atyr Pharma Says Resolaris Preliminary ...

Entity Recognition, Entity Disambiguation and Entity Salience

TechAtBloomberg.com

Bloomberg

Engineering

Challenges

Various types of documents processed

- News from hundreds of sources, including social media
- Equity research, transcripts and filings

Different Requirements

- High precision vs. High recall
- News editing vs. Search vs. News tagging

Financial entity types

- Products, currencies, etc.

Latency constraints

… all these present opportunities for research.

Bloomberg

Engineering

Temporal Drift

Temporally-Informed Analysis of Named Entity Recognition

Shruti Rijhwani, Daniel Preoțiuc-Pietro – ACL 2020 [video]

TechAtBloomberg.com

Bloomberg

Engineering

Motivation

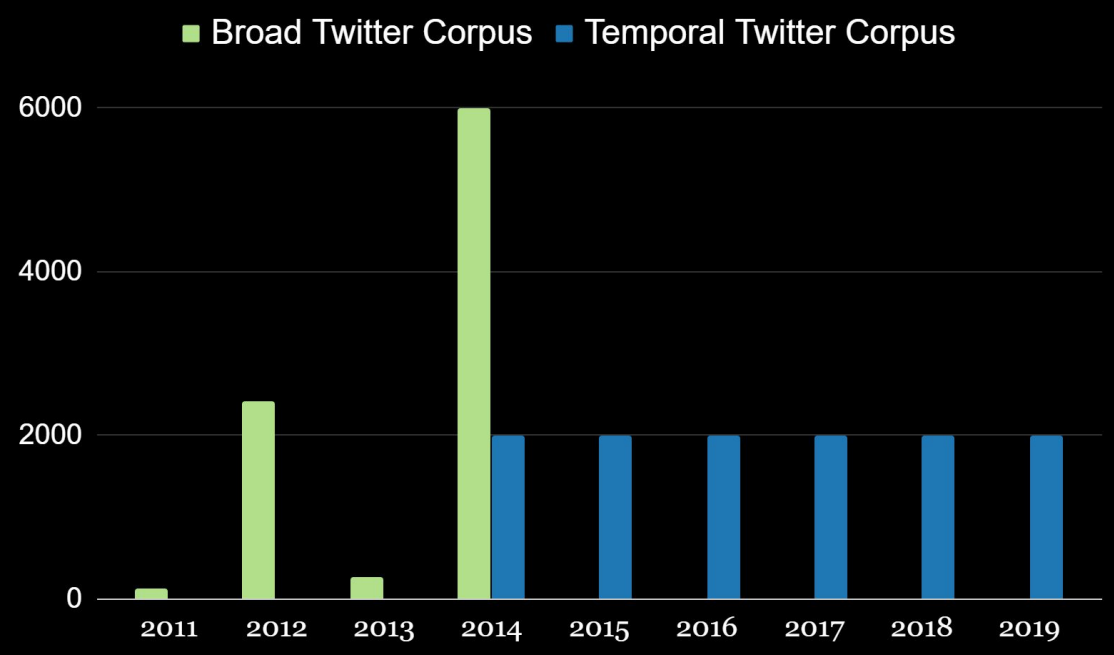
- Text data evolves over time because of **changes in language use**
- The usual setup of large-scale NLP models:
  - Trained and evaluated on **random splits of available data**
  - Make predictions on data in a **future time period**
  - Does not take temporal data drift into account
  - Lower performance on future data compared to the test set
- **Temporal information in modeling** may lead to better performance

Research Questions

- **Does temporal drift in training data affect performance?**
- **Can we leverage the temporal information of the training instances to improve performance?**
- Case Study: Named Entity Recognition on English Twitter Data
  - Readily accessible timestamp information
  - Users on social media post about current events
  - Reflects changes in language use faster than other sources of data

Temporal Twitter Dataset

- Existing Twitter NER datasets do not have sufficient temporal diversity
  - Broad Twitter Corpus (2016)
  - Data from 2009-2014, unevenly distributed
- Temporal Twitter Dataset
  - 2,000 tweets from each year between 2014-2019
  - Sampled using the same strategy as the Broad Twitter Corpus
    - Six English-speaking locales
    - Twitterati, i.e., individuals from array of domains including musicians, journalists and celebrities
    - Mainstream news organizations, both larger networks and local news outlets

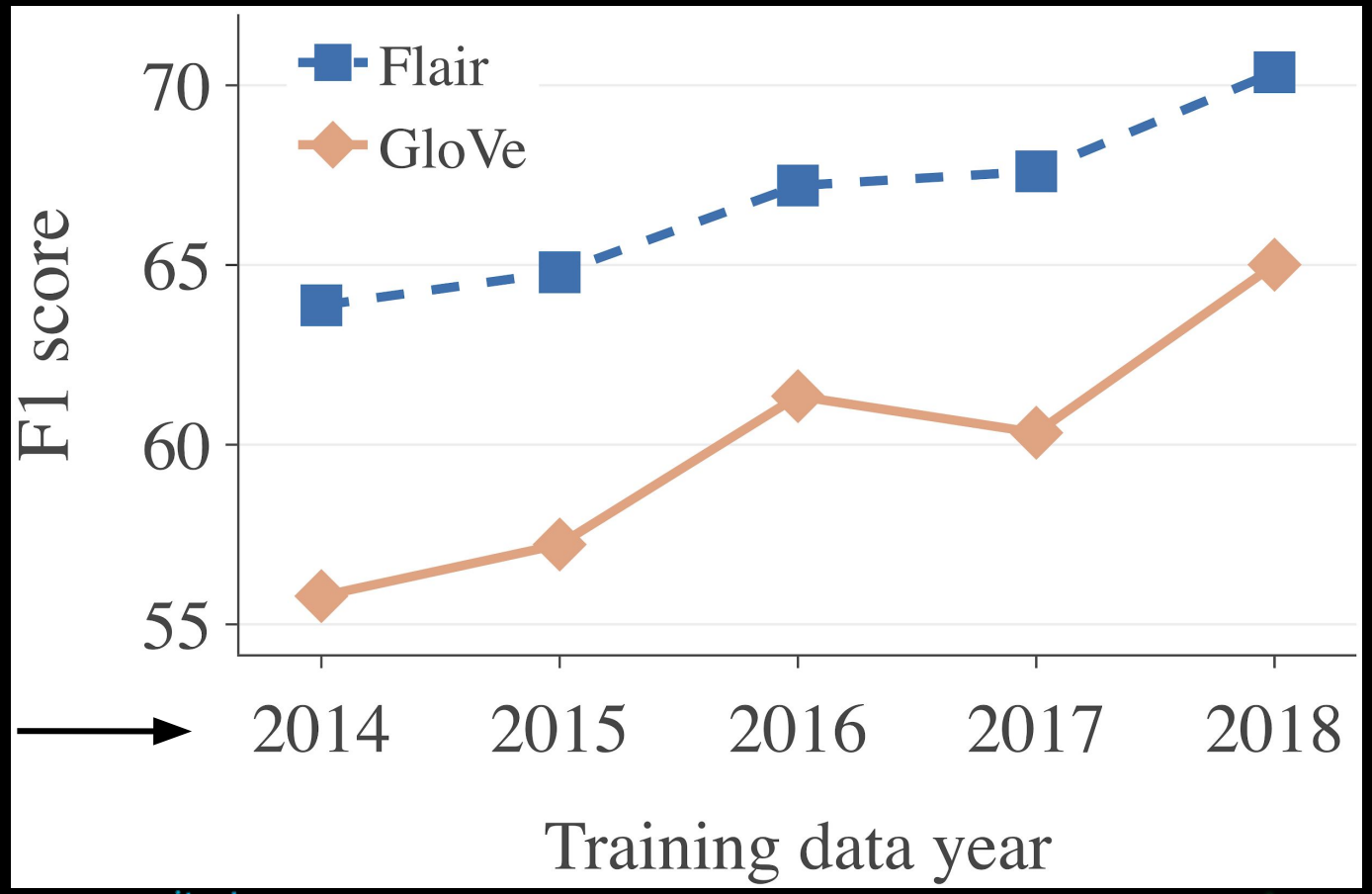


Experimental Setup

- NER Model Architecture
  - **Character and word embeddings** to represent the input text
  - **Bi-LSTM** to encode the input text
  - **CRF** to make a globally normalized prediction
- Word Embeddings
  - **GloVe:** static word embeddings
  - **Flair:** contextual word embeddings
  - All trained on data from before 2014
- Data splits
  - Train on 2014-2018 tweets
  - Validation and test on random splits of 2019 tweets
  - **Simulates a "future time period" for inference**

Does Temporal Drift in Training Data Affect Performance?

Data temporally closer to the target data gives better performance

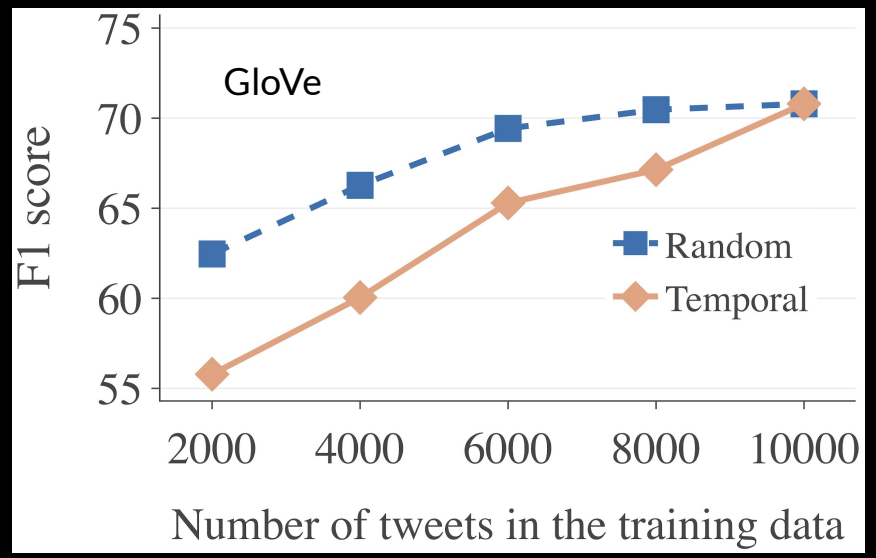


Trained on each year individually
Each model has the same number of examples

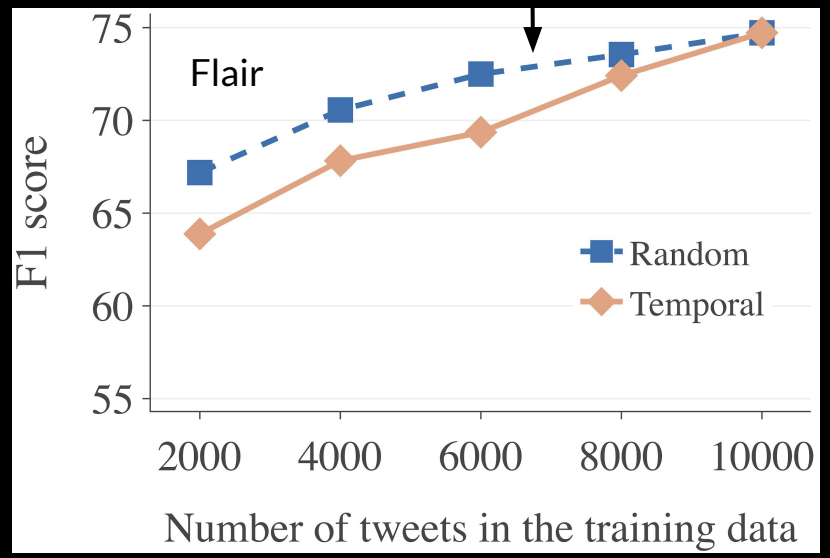
Does Temporal Drift in Training Data Affect Performance?

Temporal distribution of the training data impacts performance

- **Random:** Sample randomly from all years
- **Temporal:** Cumulatively add data in sequence of years

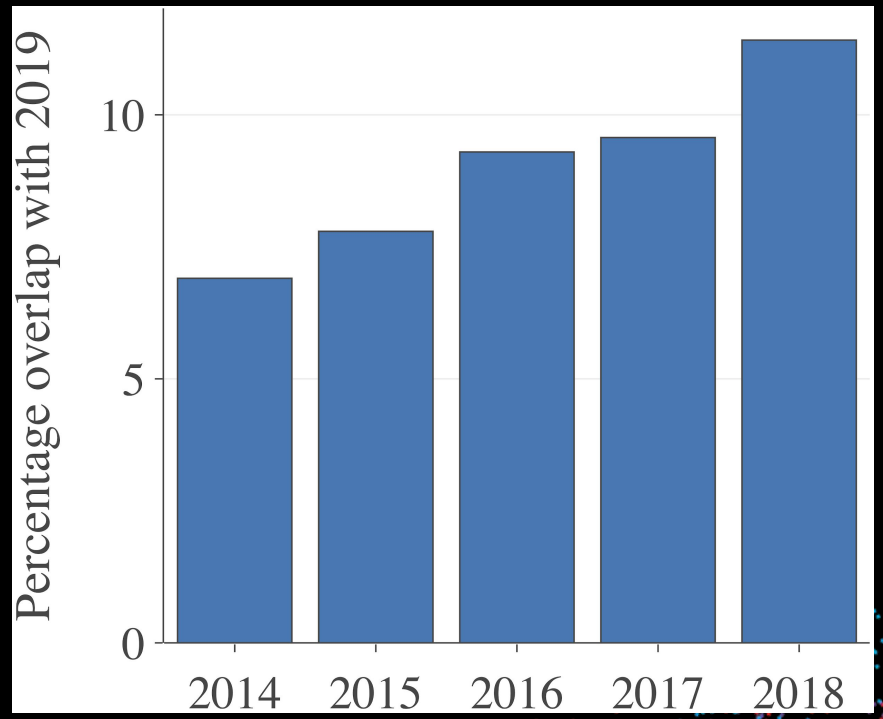


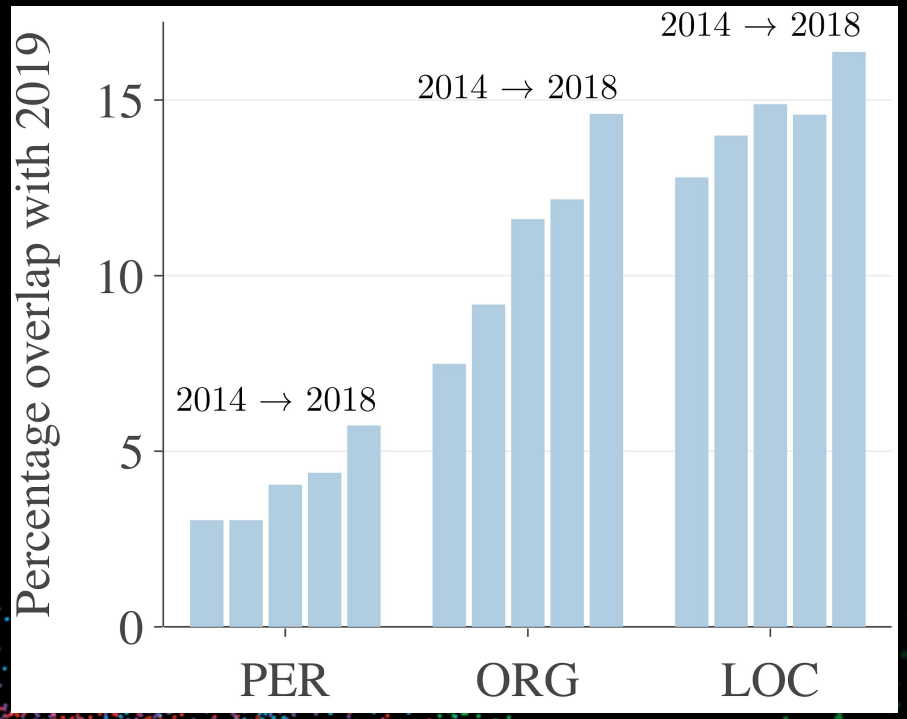
Data sampled from all years has better performance



Analysis: Entity Mention Overlap

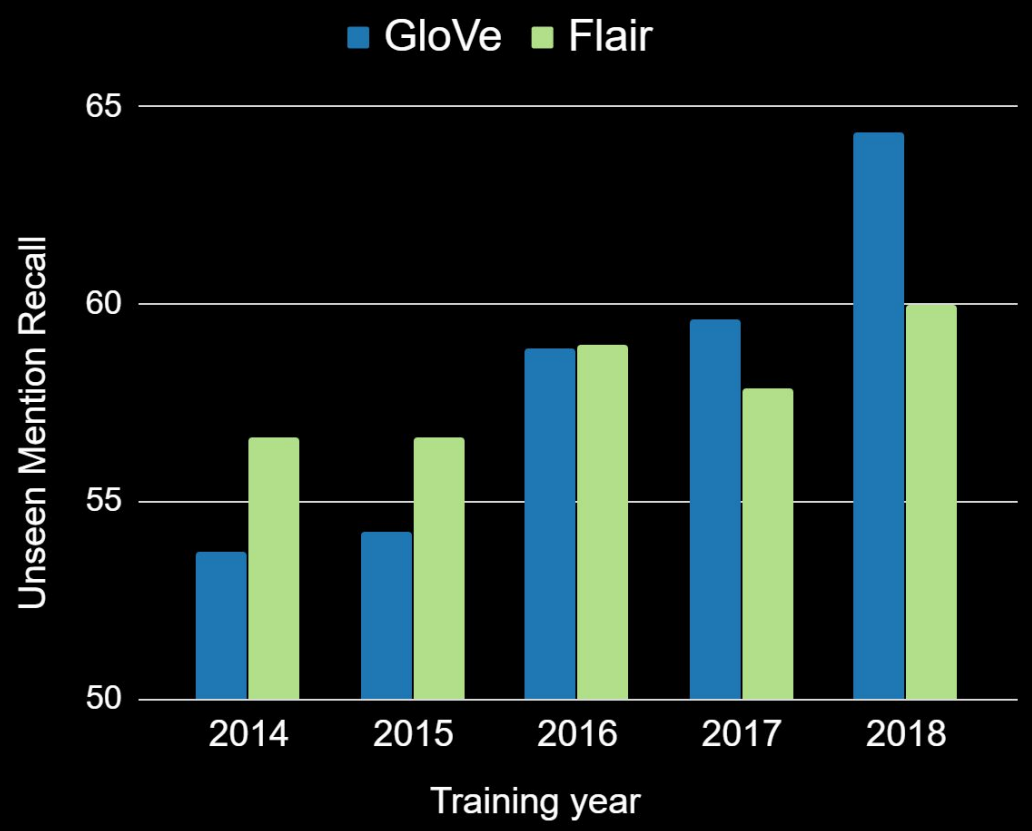
- One potential reason for better performance is the overlap of entity mentions between the training and test data
- Overlap increases as we get temporally closer to the target data: **overall and type-wise**





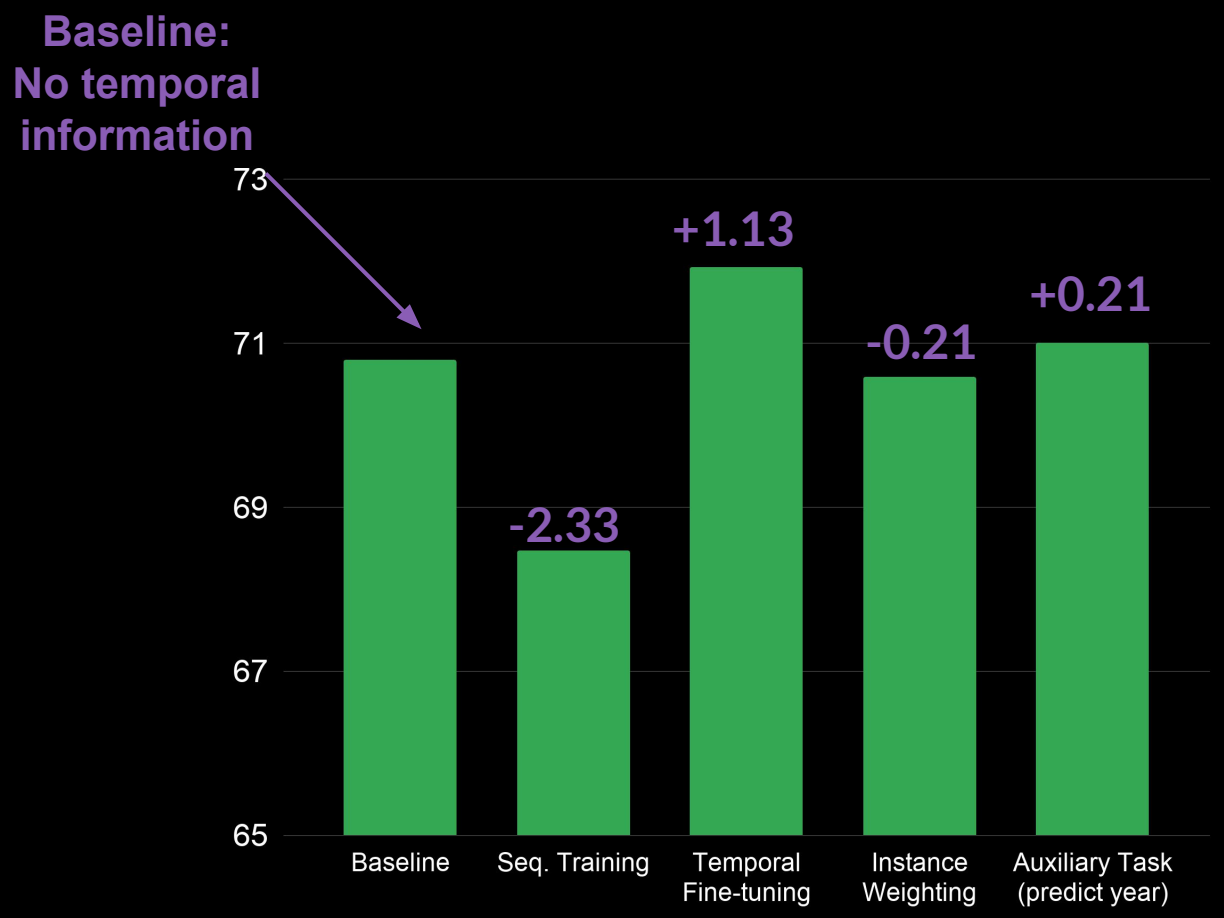
Analysis: Model Performance

- Is surface-level overlap the only factor in temporal drift?
- **Recall for mentions unseen in the training data** increases as we move temporally closer.
- The model is able to learn **more relevant context.**



Modeling Temporal Information

- Initial exploration of methods
  - Do not require significant modifications to the model
- **Sequential training:** Train the model on each year sequentially
- **Temporal fine-tuning:** Train the model on data from all years and fine-tune on 2018
- **Instance weighting:** Double the weight of training data from 2018
- **Year prediction:** As an auxiliary task, use shared representations to predict the year of the instance during training



Experiments with GloVe embeddings

Takeaways

- Temporal information is useful and can be leveraged in training
  - Annotated data is collected across a time range
- Temporal information is useful and can be leveraged in training
  - Training on data from a closer time period to the target leads to better performance
  - Fine-tuning models on temporally close data improves performance over simply combining all data for training
- Get our data set at https://zenodo.org/record/3899040

Multi-Domain Modelling

Multi-Domain Named Entity Recognition with Genre-Aware and Agnostic Inference

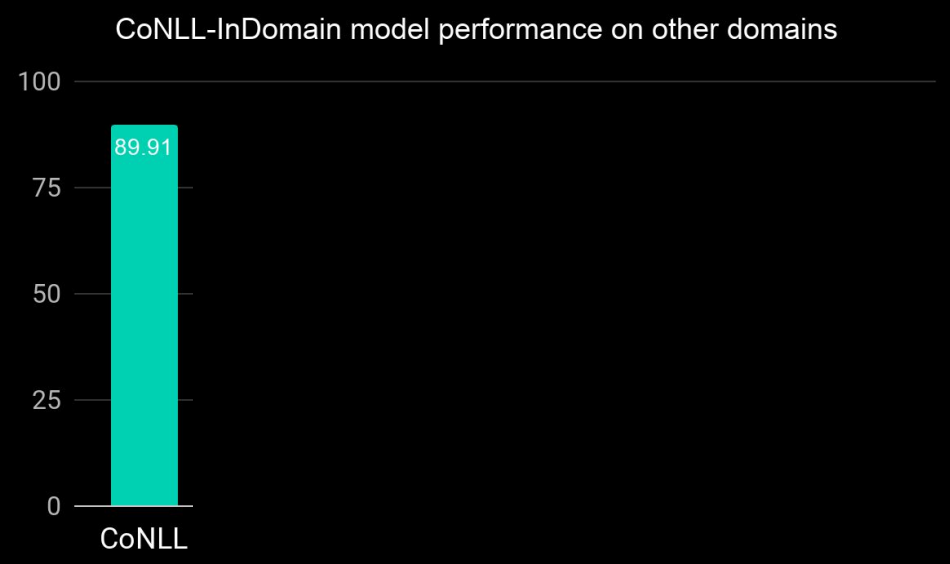
Jing Wang, Mayank Kulkarni, Daniel Preoțiuc-Pietro – ACL 2020
[*video*]

TechAtBloomberg.com

Bloomberg
Engineering

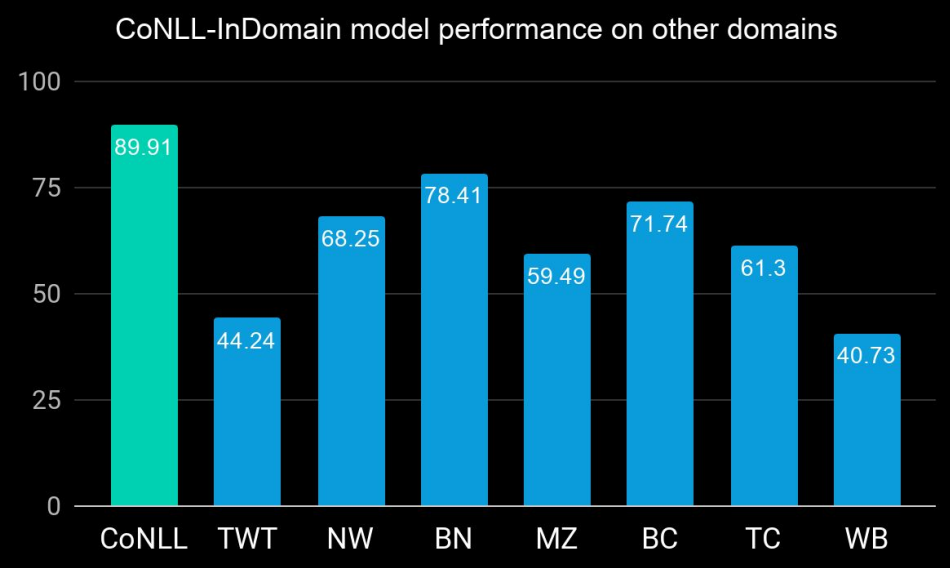
Motivation

- NER models achieve high performance when tested on data from same domain



Motivation

- NER models achieve high performance when tested on data from same domain
- … but generalize poorly on data from other domains

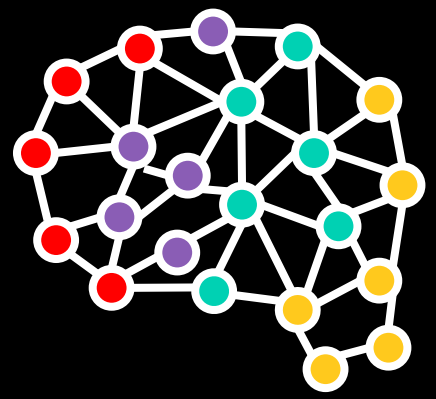


Multi-Domain NER

Ideally, one model should be able to generalize to multiple domains

- This problem is called **domain adaptation**
- Traditionally, this is posed as single-source, single-target
- Transitioning to multiple-target domains is non-trivial

ONE MODEL TO



RULE THEM ALL!

Benefits of Multi-Domain NER

- Leverage commonalities across multiple domains
- Leverage specific information for a single domain
- Better generalization to other domains
- Simplifies model maintenance

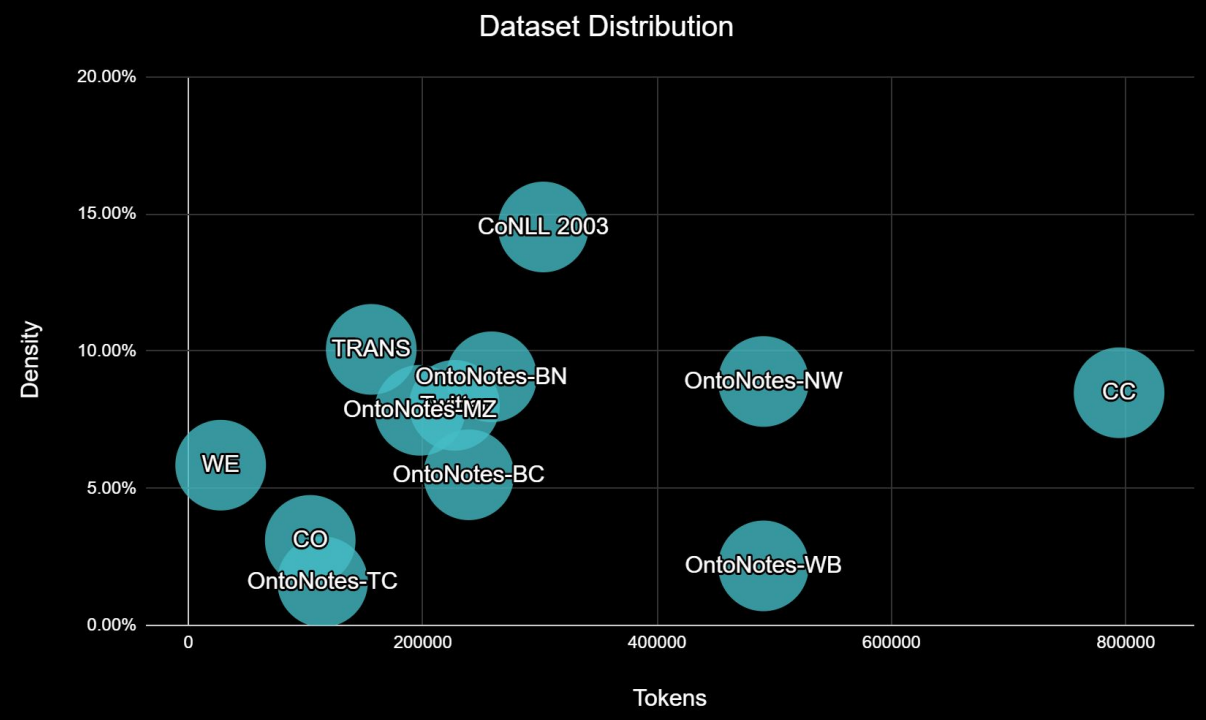
… when compared to training independent models

Experimental Setups

- Multi-domain with known domain information
  - User provides data from a domain used in training, and knows the domain label
- Multi-domain with unknown domain information
  - User provides data from a domain used in training, but does not know the domain label
- Zero-Shot domain
  - User provides data from a completely different domain (i.e. not in training)

Datasets

- **CoNLL 2003** - News Articles
- **Twitter** - 22K English Tweets
  - Temporally-informed Analysis of Named Entity Recognition (ACL 2020)
- **OntoNotes** (6 genres)
  - **NW** - newswire
  - **BN** - broadcast news
  - **MZ** - magazine
  - **BC** - broadcast conversation
  - **TC** - telephone conversation
  - **WB** - web data
- **Zero-shot documents** (4 genres)*
  - **Zero-shot-A**
  - **Zero-shot-B**
  - **Zero-shot-C**
  - **Zero-shot-D**



Bloomberg
Engineering

* Data only used in testing

Base Model Architecture

- **BiLSTM-CRF** [Lample et al., 2016]
  - Word Embeddings: GloVe + FastText
  - Character Embeddings: Randomly Initialized

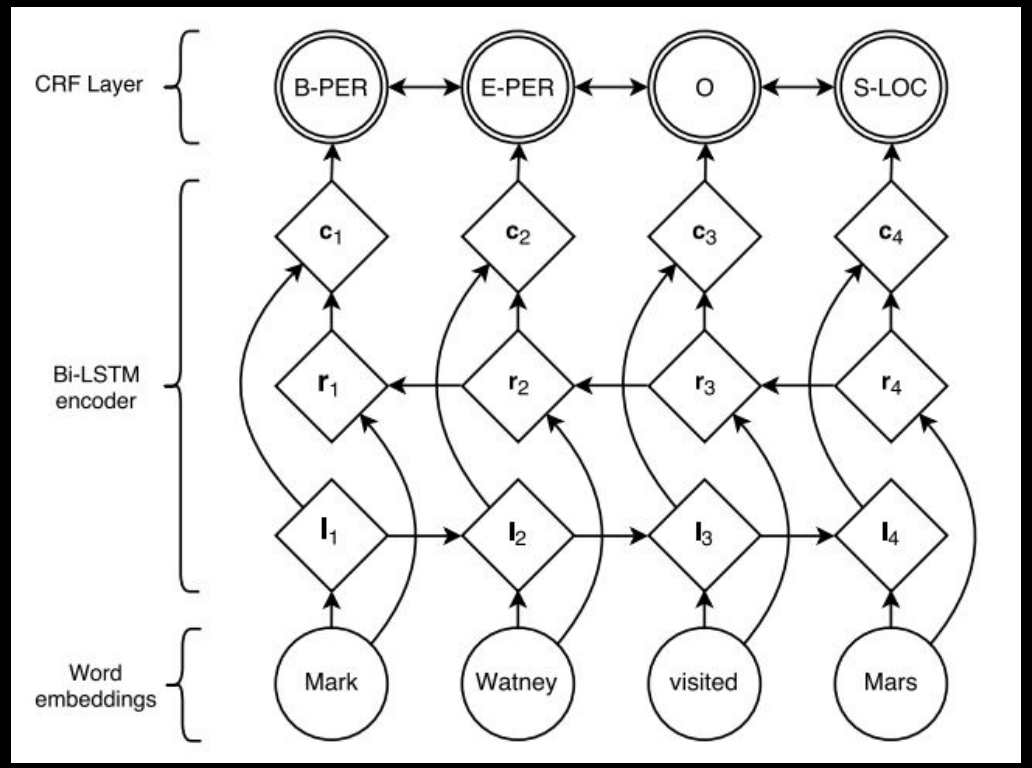


Figure from "Neural Architectures for Named Entity Recognition"

Baseline Models - InDomain

- Train one NER model for each domain
- Methods:
  - InDomain
  - InDomain + DomainClassifier

    - Train a domain classifier that will determine what in-domain model to run
- Drawbacks:
  - Poor generalization
  - Overhead of add new genres

Baseline Models - PoolDomain

- Global model by pooling data from all genres
- Methods:
  - PoolDomain
  - PoolDomain with INIT Strategy: Initialize weights from learned domain and fine-tune
  - PoolDomain with Gradient Reversal Layer
  - PoolDomain by learning domain embedding with word embeddings
- Drawbacks:
  - Discard genre-specific information

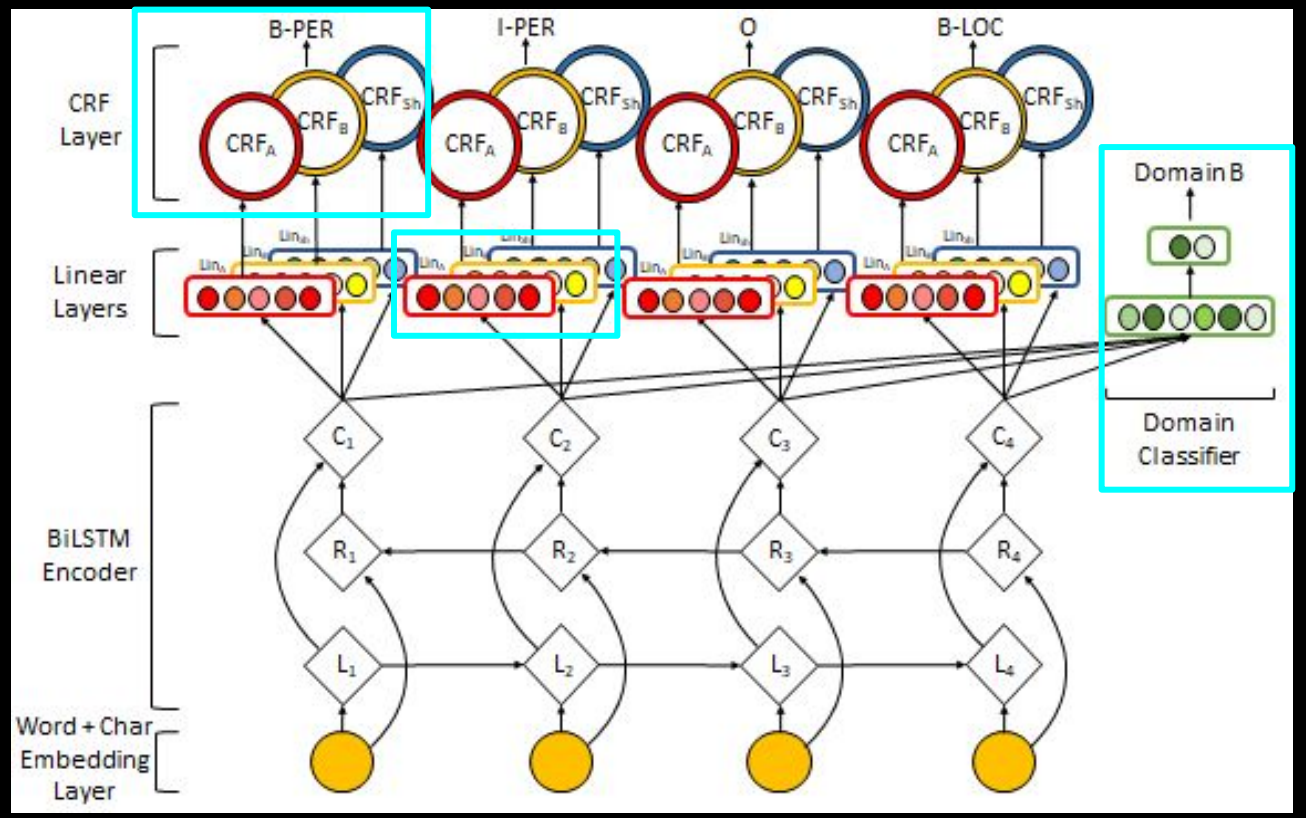
Proposed Model: MultDomain-SP+Aux Domain Learning

Add three components to the base architecture

+ domain shared (S) and private (P) CRFs

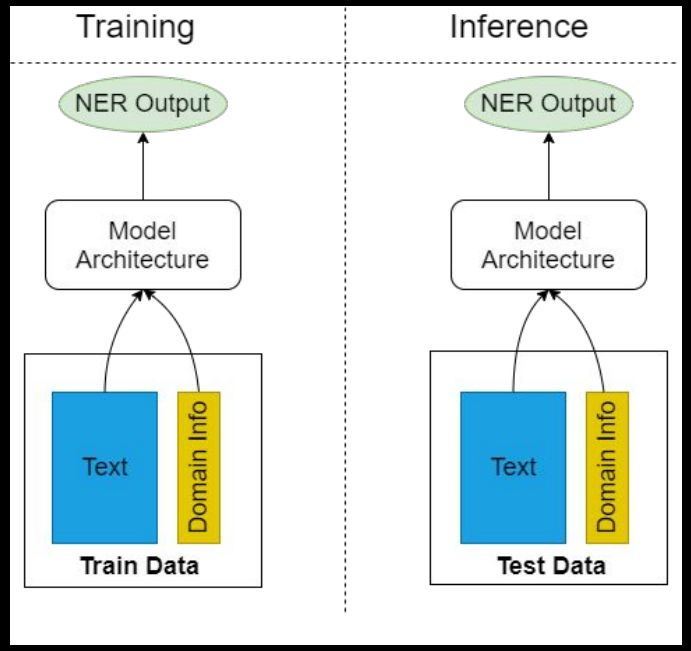
+ domain shared (S) and private (P) linear layers

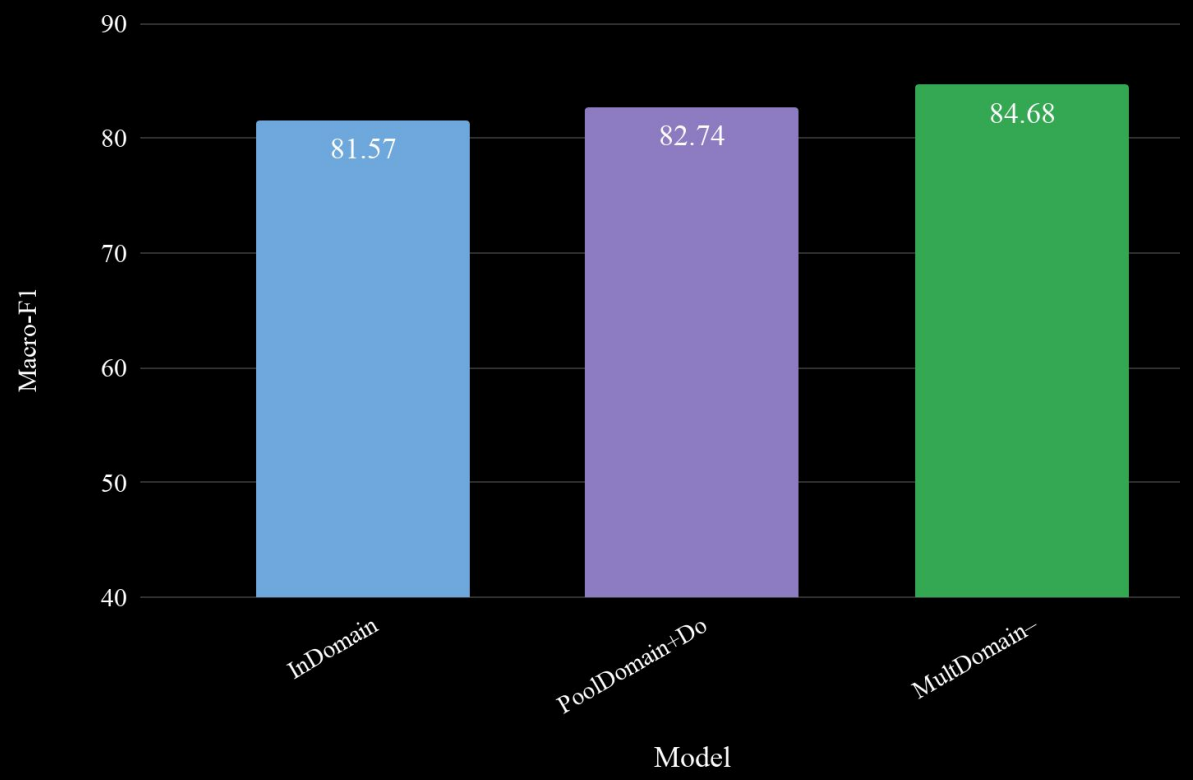
+ auxiliary task (Aux) of domain classification



Experimental Setup #1

- Multi-domain with known domain information
  - User provides data from a domain used in training, and knows the domain label

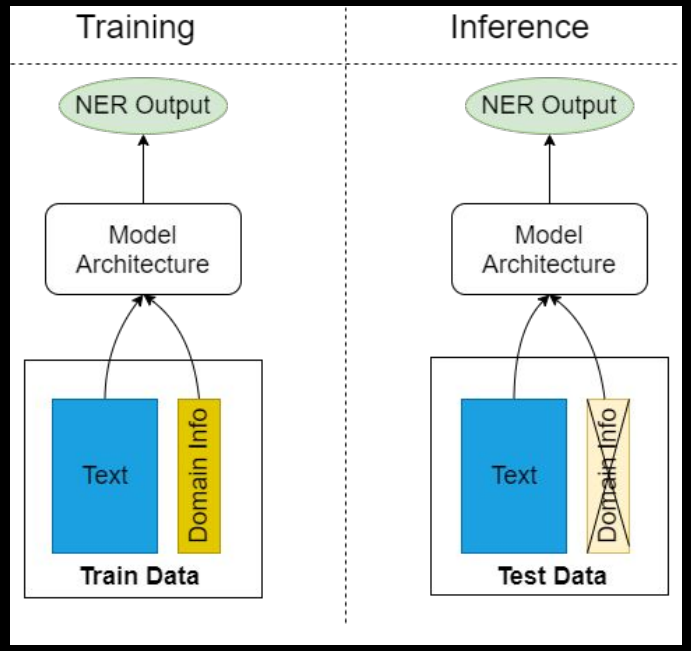


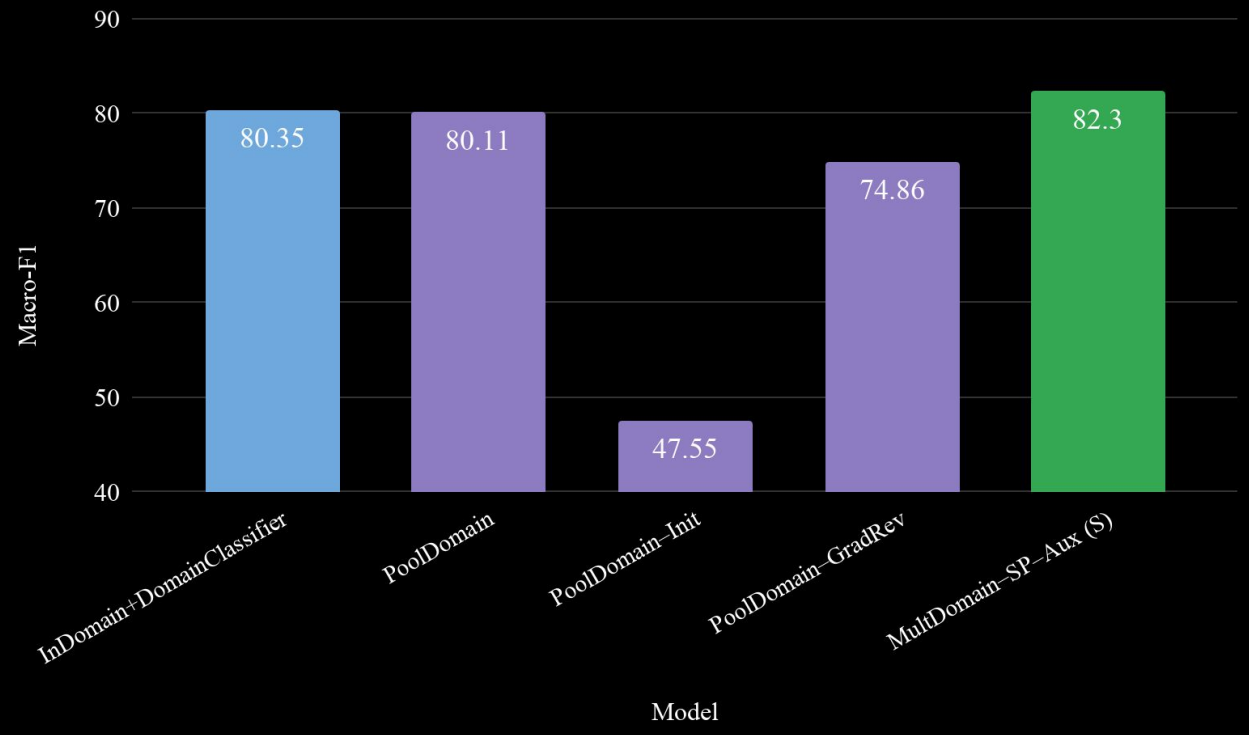


Bloomberg
Engineering

Experimental Setup #2

- Multi-domain with unknown domain information
  - User provides data from a domain used in training, but does not know the domain label

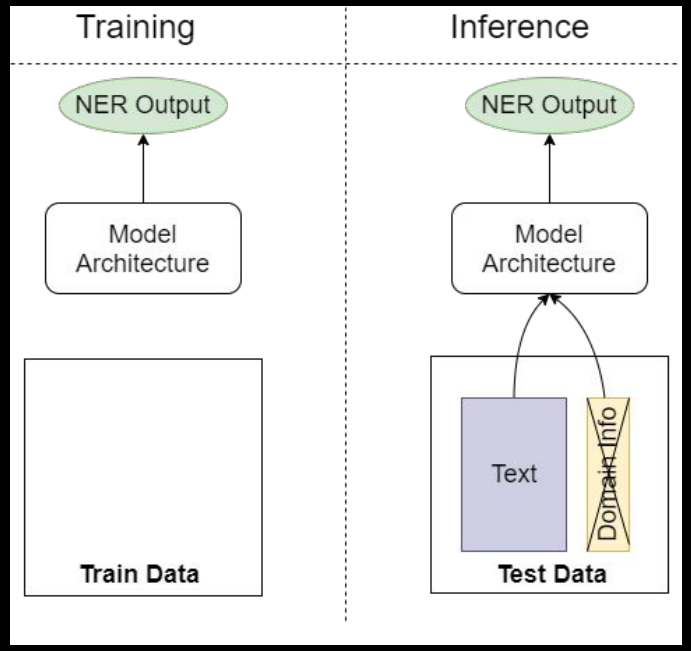


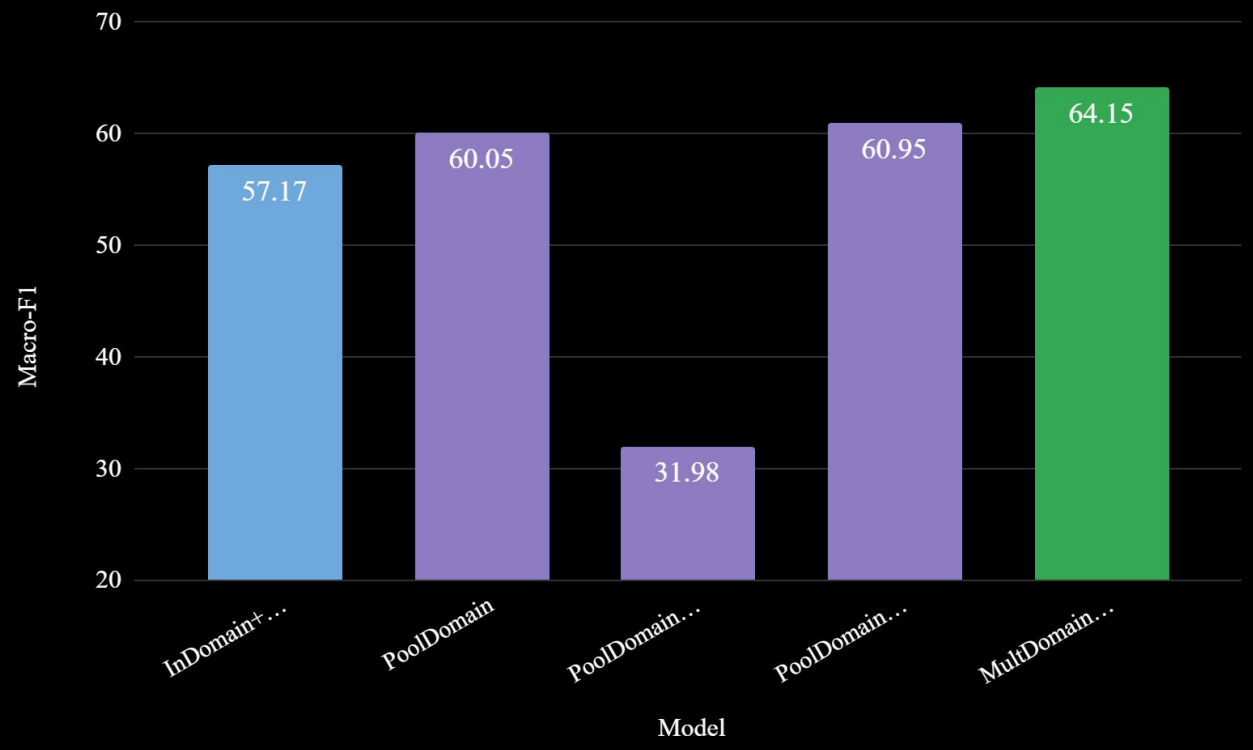


Bloomberg
Engineering

Experimental Setup #3

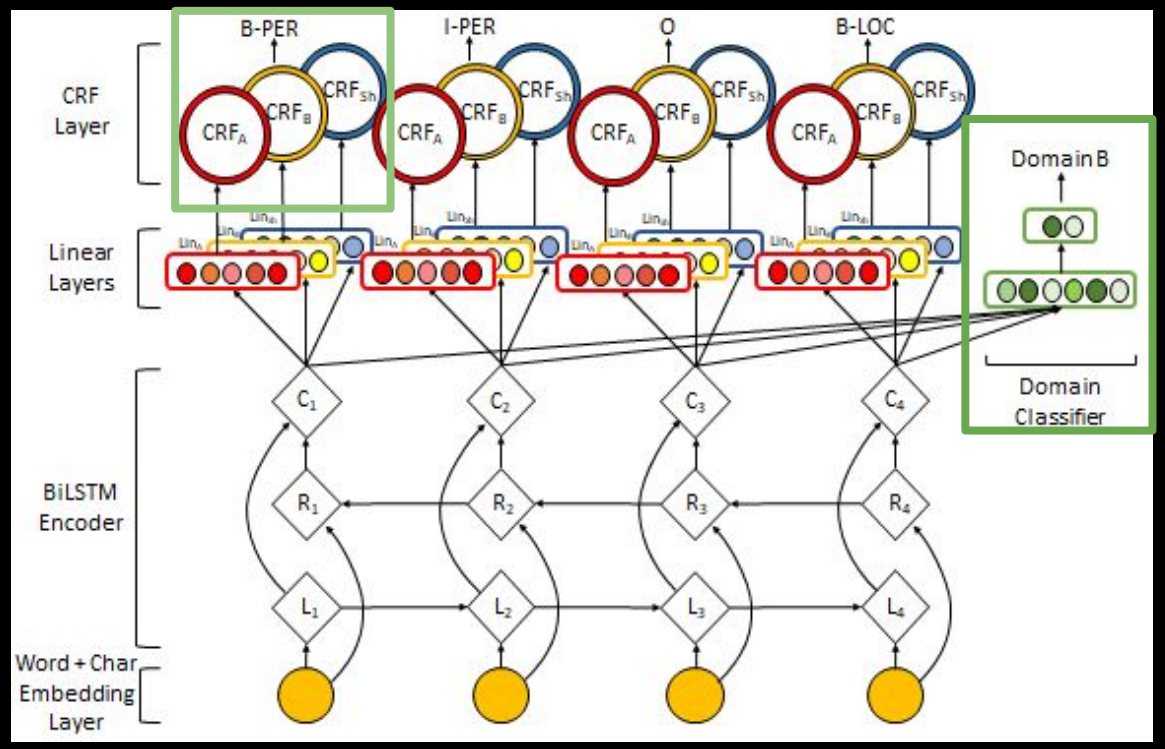
- Zero-Shot domain
  - User provides data from a completely different domain (i.e. not in training)

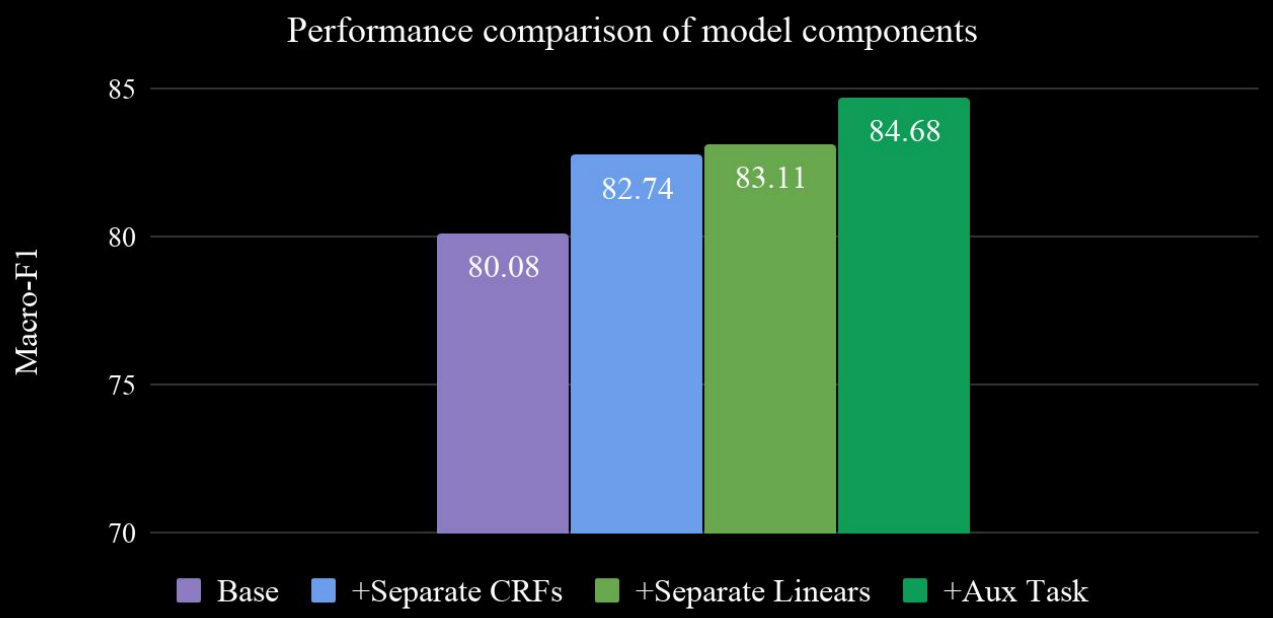




Bloomberg
Engineering

Component Ablation Study





Takeaways

- Multi-Domain NER is practically important
- Experiments using three real-world scenarios
- Proposed a robust NER model that works across multiple domains
  - Up to **+5 F1** compared to baseline approaches

High Precision NER

A Semi-Markov Structured Support Vector Machine Model for High-Precision Named Entity Recognition

Ravneet Arora, Chen-Tse Tsai, Ketevan Tsereteli, Anju Kambadur, Yi Yang – ACL 2019 [video]

TechAtBloomberg.com

Bloomberg

Engineering

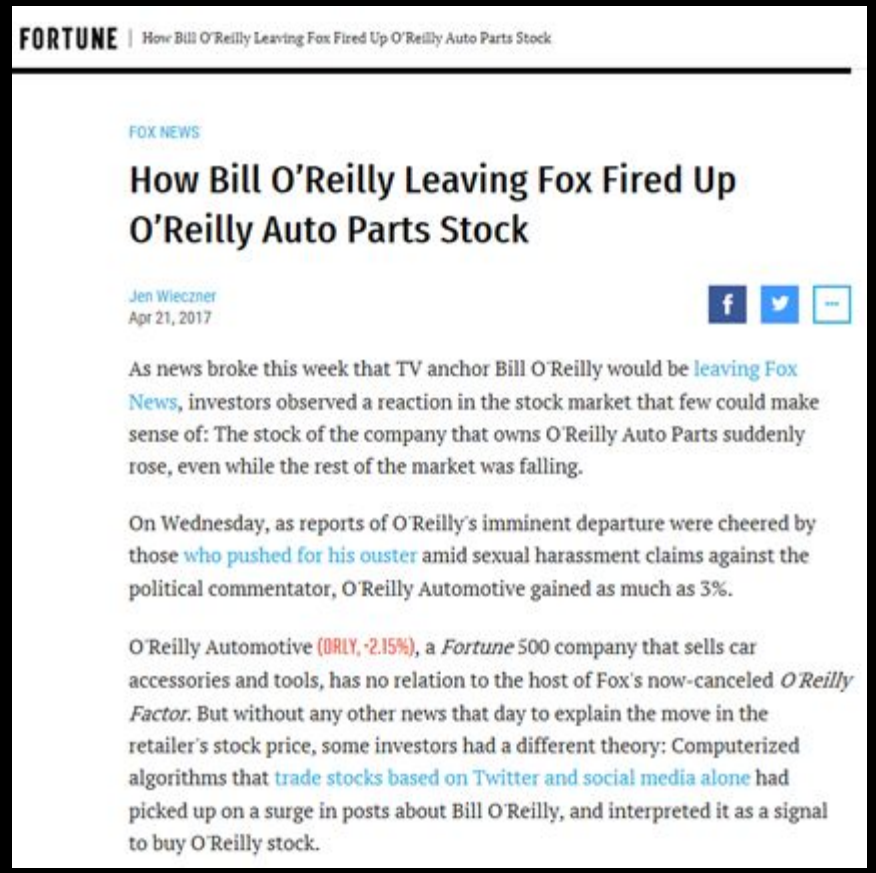
High Precision

Task: High precision NER

Focusing on a specific entity type (e.g. ORG)

Many real-world applications demand high precision:

- Finance
- Bio-medical



FORTUNE | How Bill O'Reilly Leaving Fox Fired Up O'Reilly Auto Parts Stock

FOX NEWS

How Bill O'Reilly Leaving Fox Fired Up O'Reilly Auto Parts Stock

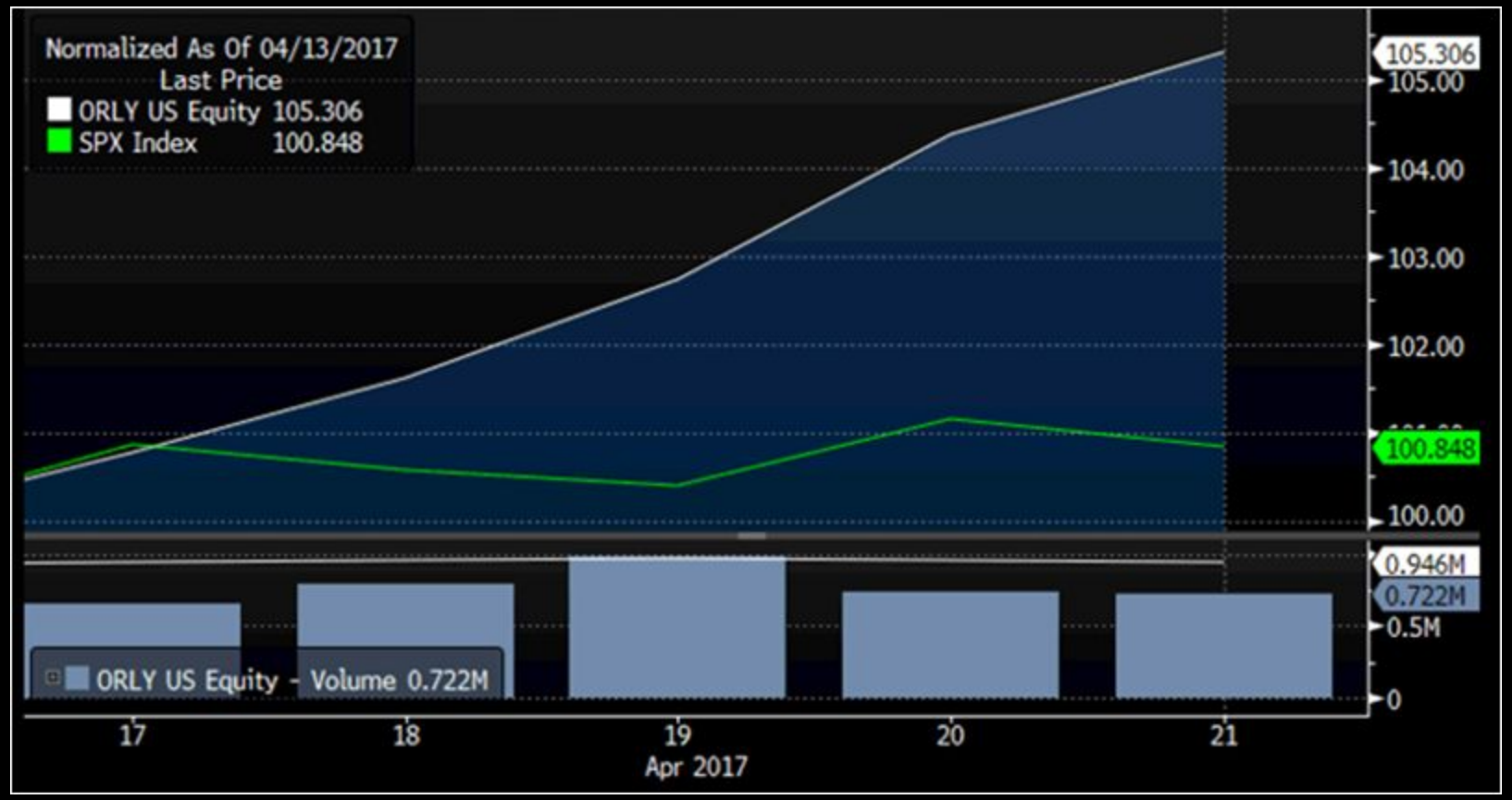
Jen Wieczner
Apr 21, 2017

As news broke this week that TV anchor Bill O'Reilly would be leaving Fox News, investors observed a reaction in the stock market that few could make sense of: The stock of the company that owns O'Reilly Auto Parts suddenly rose, even while the rest of the market was falling.

On Wednesday, as reports of O'Reilly's imminent departure were cheered by those who pushed for his ouster amid sexual harassment claims against the political commentator, O'Reilly Automotive gained as much as 3%.

O'Reilly Automotive (ORLY, -2.15%), a *Fortune* 500 company that sells car accessories and tools, has no relation to the host of Fox's now-canceled *O'Reilly Factor*. But without any other news that day to explain the move in the retailer's stock price, some investors had a different theory: Computerized algorithms that trade stocks based on Twitter and social media alone had picked up on a surge in posts about Bill O'Reilly, and interpreted it as a signal to buy O'Reilly stock.

High Precision



TechAtBloomberg.com

Bloomberg

Engineering

High Precision

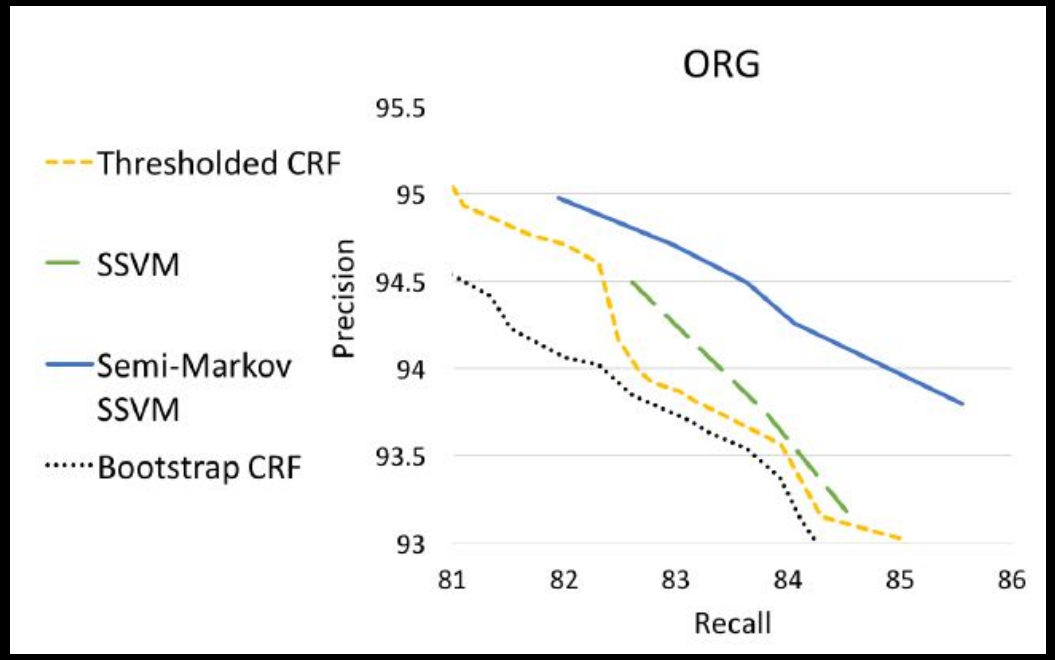
Objective:

Train models that have high precision on one type

At the expense of other types

Method:

Semi-Markov SSVM model outperforms simple thresholding



Takeaways

Named entity detection is central to finance and Bloomberg:

- Users
- Clients

Bloomberg provides unique challenges and opportunities:

- Data sources
- Use cases
- Requirements
- Latency

Thank You!

We are hiring for full-time positions:

- NYC – https://careers.bloomberg.com/job/detail/86598

Summer 2021 research internships:

- https://careers.bloomberg.com/job/detail/84208

For more information about our work, research and academic outreach programs:
https://www.TechAtBloomberg.com/ai/

TechAtBloomberg.com

Bloomberg

Engineering